

DOCUMENT RESUME

ED 163 017

TM 007 254

AUTHOR Childs, Roy
TITLE Item Banking. Basic Testing Series.
INSTITUTION National Foundation for Education Research in England
and Wales Publishing Co., Windsor, Berkshire,
(England).
NOTE 14p.; Contains occasional small print
EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Foreign Countries; Item Analysis;
*Item Banks; Item Sampling; Secondary Education;
*Test Items
IDENTIFIERS Great Britain; *Rasch Model; Rasch Scaled Scores

ABSTRACT

This pamphlet describes the exciting potential of item banking--a new approach to testing which combines both comparability of scores with flexibility of test format. Item banks are collections of items where the characteristics of each item is known and these characteristics can be summed to describe a test made from such items. The principle behind the calibration of such items is explained in relation to the Rasch model. The use of item banks provides more flexible and appropriate tests than traditional standardized tests. Although in theory the Rasch model suggests exciting possibilities for the future of achievement testing, there are still some practical problems to be solved. (Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Basic Testing Series

ITEM BANKING

by Roy Childs
Psychologist



NFER Publishing Company

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. Field
NFER Pub Co.

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

ED163017

TM007 254

DISTRIBUTED by the
INSTITUTE OF PSYCHOLOGICAL RESEARCH, INC.
34 FLEURY STREET WEST, MONTREAL, QUEBEC
H3L 1S9 TEL (514) 382-3000

Introduction

One of the main interests that teachers will have in the subject of item* banking is to do with their freedom. The diversity of approach to curriculum subjects displayed in this country is something which the teaching profession has long treasured and it would not be easy, even if it were desirable, to reduce the involvement of the teacher in the process of deciding what to teach children and how to do it. It is partly this desire to be involved in curriculum matters which has made a lot of teachers suspicious of educational assessment.

There can be no doubt that public examinations such as GCE, O- and A-levels exert considerable influence on the curriculum and many teachers believe that widespread use of objective tests could do the same. The impetus behind the development of the CSE mode 3 exams was specifically aimed at overcoming this constriction of the curriculum by the need to evaluate how it is being learnt. This has led to the problem of achieving comparable results between a wide variety of exams. As a result many outside bodies still do not recognise mode 3 exam grades.

It has seemed impossible for assessment and curriculum freedom to exist alongside each other and yet there is increasing pressure for more assessment and more teacher accountability stemming from the controversial debate concerning educational standards. The fears within the teaching profession that more assessment will lead to a narrowing of the curriculum to only those aspects which can be assessed, together with a lesser involvement of the teacher in curriculum matters, are now very real.

Item banks, however, provide the hope that this will not happen; that assessment can be flexible enough to cope with the great diversity of our educational system and still provide results which are meaningful, unambiguous and comparable. Every teacher concerned with his or her future role in education should have a considerable interest in the present development of Item Banks.

What are Item Banks?

Before we can demonstrate how item banks can be used to overcome some of the problems mentioned above, we must be clear what we mean by the term "item banks". The term is often used for any collection of items which do not form a specific test. This is not

* The term "item" is used for each individual question or unit which can be marked pass or fail. A test is therefore made up of a number of items.

how we intend to use the term. Consider the following examples: Examination Boards have access to large numbers of previous questions which can form the basis of a new exam; professional test constructors such as those at the NFER have access to large numbers of items which can be used to construct a variety of tests designed for different purposes; teachers themselves, will construct tests from the questions or items which were used in a previous year's testing. The term 'item bank' will not be used to describe collections of items such as these. This is because a test which is constructed by taking a subsample of items from such a collection cannot be used to compare performance on *another* test which was made up from a different sample of items designed to reflect a different curriculum. A score obtained on either test could reflect the test length or the difficulty of the items, as much as the ability of the particular child taking the test. We have no idea how each individual item contributes to the total assessment of the child, because each item is essentially an unknown quantity.

Traditionally constructed tests get over this problem by trying out the new test on a large sample of children. Bad items can then be 'weeded out' according to well defined statistical principles which help ensure that a test is composed of items which all measure the same kind of area or trait. However, interpretation of the test score is then based on the test as a complete unit. The individual items are "locked" into the test such that their contribution to any objective assessment is in their contribution to the total test score.* If any of the items are removed or changed, the effect on the test as a whole is unknown. Hence traditional tests cannot be tampered with to provide a better reflection of a particular syllabus without altering the interpretation which can be made from the test score.

If, however, the contribution of each item were recorded separately in such a way that the characteristics of a test were obtained by summing the characteristics of the individual items, then we would have a system whereby we would be at liberty to add or to remove items as freely as we saw fit. Such a test would still enable us to make meaningful comparisons. It is the calibration** of

* A teacher is obviously at liberty to examine which kinds of items the pupil gets right or wrong in order to make a more subjective judgement of the child's problem. We are, however, talking about limitations of *objective* measurement which requires that we interpret and generalise from objectively obtained scores.

** The labelling of items statistically is called "calibration". The main statistical characteristic in the *Secondary Maths Item Bank* currently published by the NFER is the item's difficulty which is discussed later.

individual items in just such a way which we will use to define an item bank. A group of uncalibrated items, such as those built up by teachers and such as those usually available to examination boards we will call item pools.

A simple analogy may help to demonstrate the difference between the two. Consider two boxes, one which contains a number of ordinary thermometers and another which contains a number of blank thermometers (i.e. where the temperature scale has been erased). The second box is like our item pool. Any thermometer taken from this box is an unknown quantity. We do not know if it is appropriate for the range of temperatures we are to measure. If it does happen to cover the appropriate range of temperature we can use it to see whether one cup of water is hotter or colder than another, but we cannot say *how much* hotter one cup is than another. Neither can we use two blank thermometers to compare the temperature of two cups of water separately.

This is because the level reached on each thermometer cannot be compared. Similarly, a test made up of items from an item pool, is also an unknown quantity and may or may not be appropriate for the ability level concerned. A score obtained on another test made up from a different selection of items (from the same item pool) cannot be used to compare the performance of different children taking the first test.

Our item bank is more equivalent to the other box of ordinary thermometers. Tests constructed from the bank can be used to obtain a scale score. Comparisons between tests becomes equivalent to comparing the temperature of two thermometers which have the same scale engraved on the side.

This analogy describes what the item banks could achieve. Further details on how this could be achieved are given in the section "A Measurement Scale for Item Banks". However, the next section gives more details on the practical use of Item Banks.

How Would We Use an Item Bank?

Before describing the practical implications of using an item bank it is necessary for the reader to understand some basic characteristics of items used to stock an item bank. There are two major constraints to which items in the bank should conform. Both these constraints are related to the concept of *independence*.

The first constraint can be called the 'independence of items from each other' which requires that each item must be able to stand on

its own. In other words the contribution that each item makes to any test must not be affected by the other items in the test. This can be checked by observing any change in the items' characteristics* when in the company of another set of items. It is sometimes found, for instance, that an item is easier if certain other items are attempted beforehand. However, the consequence of stocking the bank only with independent items is that they can be taken from the bank in any order and put alongside any other items for the purpose of constructing the test.

The second requirement can be called the "independence of items from the people taking it". This means that the essential characteristics of the items must not change when given to different pupils. Returning to our analogy, it would be rather difficult to compare the temperature of a cup of water with the temperature of a cup of milk if our thermometer was sensitive to the substance in the cup as well as being sensitive to the actual temperature. The result would be a scale which fluctuates for either of these two variables and a single measurement could not tell us how much is due to one variable and how much is due to the other.

The consequence of these two kinds of independence is a flexibility whereby the bank can be used to construct any number of tests to any kind of specification for any group of people as long as the bank contains suitable items. Since the items have all been calibrated, it becomes a simple procedure to combine items and obtain a test which has the desired content *as well as* the desired psychometric properties. In other words the test user can become the test constructor without the disadvantages of traditional teacher-made tests (see reference 1). Teachers could define those areas of their syllabus which they wanted to assess together with the level of difficulty they considered most appropriate for their pupils, and suitable items could be drawn from the bank to form a test. How this is actually achieved in practice depends on a number of factors.

The experimental item bank currently produced by the NFER (the *Secondary Maths Item Bank* which covers the ages 12+ to 14+) is available in the form of a booklet which contains 185 items covering a number of subject areas in mathematics. Teachers can buy the whole bank and the items for any specified test are selected by the teacher. However, this is likely to be impractical if item banks are

* The characteristics referred to here are the statistical characteristics which, when added together, give the statistical character of the test.

to contain the large numbers of items necessary to cover large subject areas such as mathematics or language. It is anticipated that, in the future, item bank facilities rather than the item bank itself will be made available to the user. To use the facility the teacher would simply submit a specification of those areas he would like assessed and items would be selected which fit the specifications. These would then be printed to form a test of the required length and difficulty.

There remain two questions which the user of traditional tests is likely to ask. The first is "How do we interpret scores on such a 'made to measure' test?" The other is "What is the reliability of such a test?" This requires the development of a measuring system for item banks.

A Measurement System for Item Banks

The previous paragraphs have assumed that the statistical characteristics of the *items* can be obtained to give us the statistical characteristics of the *test*. It was further implied that such a test would yield interpretations and generalisations at least as valid as those obtained on a traditionally standardized test. This section attempts to explain how this could be done. There are a number of statistical models of differing complexity which could be used to achieve this but the present account describes *some* of the principles behind one particular model called the Rasch Model. This is because the only item bank at present available from the NFER to teachers is the *Secondary Maths Item Bank* which used the Rasch Model for the calibration of its items.

The basic assumption of the Rasch Model is that there are two main factors or parameters which determine whether a person will pass or fail an item. These two parameters are the ability* of the person and the difficulty of the item. Although other parameters such as fatigue can significantly affect a person's score, it has been found in practice that the parameters of ability and difficulty can explain test results adequately as long as the tests are administered sensibly. The following paragraphs will give a brief explanation of how these two parameters are calculated and then how they are used to form the basis of a measuring system.

Let us first take the parameter of *item difficulty*. This can be

* For the sake of brevity ability will be used interchangeably with a person's level of achievement. The problems are not, however, identical and this is touched on later.

estimated by observing the percentage of the people who pass each item. However, if a different group of people were to attempt any particular item we would probably get a different percentage of people passing it. This, we assume, is because the group are of a different ability. How, then, do we estimate the difficulty of an item since it appears to fluctuate depending on which group is attempting it? The answer lies in our assumption that the difficulty of an item relative to others is constant regardless of who attempts it. A consequence of this assumption, although not the only consequence, is that the *order* of difficulty of the items will remain the same even when the *level* of difficulty fluctuates. We can check this by giving the items to different groups. If the items remain in the same order of difficulty our assumption is borne out in practice. The following table represents an attempt to do this:

Table 1

	Group 1 Rank order	Group 2 Rank order	Group 3 Rank order	Theoretical Hierarchy
Item 1	1	1	1	1
Item 2	2	2	3	3
Item 3	3	3	4	4
Item 4		4	5	5
Item 5			2	2

We notice that for Group 1 and Group 2 items 1, 2 and 3 are in the same order. The items can be said to fit one of the conditions of our model i.e. that they maintain their relative difficulty.

In Group 3 the rank positions for items 2 and 3 change to become 3 and 4. However, we still observe that item 1 is harder than item 2 which is harder than item 3. We have merely obtained further information which tells us that item 5 is somewhere between item 1 and 2 in its level of difficulty. In this way we can "link in" any items to this hierarchy of difficulty such that they are all uniquely ordered*. It should be borne in mind, however, that at times it would be very difficult to observe which items were harder than others. For example, if a group of pupils of high ability attempted items 1, 2 and 3 in the table and found them very easy we may find that the per cent of pupils passing each one makes them almost

* There are, of course, precautions to be taken before doing this. It should be checked that the item does maintain its relative difficulty by obtaining results for that item from at least 2 separate groups of people. Those items that do not do this are said to be sample dependent and would not be used to stock an item bank. The question of whether an item 'fits' can be determined statistically.

indistinguishable (e.g. 85%, 83%, 82%). This demonstrates that the match between pupil ability and item difficulty will affect the extent to which the items appear to be different. Sophisticated methods are therefore required for looking at data to ascertain whether a unique hierarchy of difficulty exists. However, given a set of items for which this holds true we can look at the second parameter of the Rasch Model which is a person's ability.

The calculation of the ability parameter is done by deriving it from the set of item difficulty parameters. To do this another basic assumption is necessary which is simply that a certain level of ability is required before a person is *likely* to pass an item of certain difficulty. Another model of mental measure, called the Guttman Scaling, expresses this concept in a more extreme way. It states that a person of certain ability will fail *all* items above a certain level of difficulty. This can be demonstrated with the following diagram in which the items have been placed in their unique hierarchy of difficulty. The responses of one person Mr X, have been recorded as pass or fail for each item.

Item No.	Responses of Mr X
16	fail
15	fail
14	fail
13	fail
12	fail
11	fail
10	fail
9	fail
8	fail
7	pass ← level of ability for Mr X
6	pass
5	pass
4	pass
3	pass
2	pass
1	pass

The problems of estimating Mr X's ability according to the Guttman Model becomes the problem of finding two items which are close in difficulty but where the person passes the easier one and fails the harder one. We can then express the ability scale in the same units as the difficulty scale.

In practice this situation would never occur. Items would never maintain exactly the same relative difficulties from group to group.

Likewise, no pupil would pass *all* the easy items and fail all the hard items. However, it is possible to explain what happens in real life far better if we use a model where everything is phrased in terms of probability. In other words we require a model which states that a person will *probably* pass the easy item and *probably* fail the difficult ones. In fact it would be reasonable to expect a person with a certain ability to pass nearly all the very easy items, but that his chances of passing progressively harder items would decrease until they were almost zero.

With such a model we could define a person's ability as the point at which the probability of passing an item is 50%. This definition, is, in a sense, arbitrary since the choice of any point would leave the rank order of any group of peoples' ability estimates unchanged. Such an ability scale has no real zero point. It is similar to choosing the freezing point of water as the zero for the temperature scale. Any other point could have been chosen without affecting the comparison of a hot cup of water with a cold one.

The Rasch Model is just one of a number of possible models which incorporates the concept of probability into the Guttman Model and it also defines ability as the point at which a person's ability matches the items' difficulty*. The result is that the ability estimates are on the same scale of measurement as the difficulty of the items. The ability estimate has the same numerical value as the item for which the person has a 50% chance of getting it right.

The reader may realise that a problem still remains. How can we calculate the probability that any person has of passing an item? In practice, the person either passes or fails each item, giving us no information about his probability of passing it. We can overcome this problem if we assume that a person's estimated ability will be the same as the estimated ability of those people who obtain the *same score* on a particular test i.e. those in the same "score group". We are not saying that all these people have the same ability since many of

* The Rasch Model can be written as $\text{Prob} \{ a_{pi} = 1 \} = \frac{W(\alpha_p - \delta_i)}{1 + W(\alpha_p - \delta_i)}$

where $\text{Prob} \{ a_{pi} = 1 \}$ is the probability that a person will pass item i .
 W is a constant

α_p is the ability estimate for person p

δ_i is the difficulty estimate of item i .

where ability = difficulty $\alpha_p - \delta_i$ becomes zero

$$\text{Prob} \{ a_{pi} = 1 \} = \frac{W^0}{1 + W^0} = \frac{1}{1+1} = \frac{1}{2} \text{ or } 50\%$$

them will have got a different combination of items correct in getting this score but we are saying that the *best estimate* of their ability to be obtained from the test will be from using their total score.

We therefore need to obtain ability estimates for the score groups and not for the individuals concerned. Since only a certain percentage of any score group will have got any particular item correct, we can use this to calculate the probability that any member of the score group has of passing the item. Thus, if 80% of a score group get item X correct, then the probability of passing the item for the members of the score group is 0.8.

There are other methods of calculating the probability, based on the mathematical properties of the Rasch Model, which have greater application in the context of item banking. This is because we would not normally be required to calculate ability and difficulty estimates from the same set of data. Item difficulties would be calculated when setting up the bank and only the ability estimates would be required by future users. A property of the Rasch Model is that, for any test taken from the bank for which the item difficulties are known, we can calculate ability estimates for each score group in advance. The procedure usually requires a computer programme since a Maximum Likelihood Technique is usually used to obtain the best estimate. The interested reader is referred to reference 2 (Appendix One) for a more detailed account of the procedure involved. This reference also deals briefly with the calculations of the Standard Error of Measurement (SEM) which cannot be gone into here except to say that the closer the match between item difficulty and the person's ability, the smaller will be the SEM.

How do Item Banks compare with Traditional Tests?

The advantages of item banks stem from the way they can be used flexibly. This enables us to construct tests for specific purposes which should therefore increase the efficiency of testing. To make this point clearer, a teacher may wish to test the success of a particular course of instruction. Items can then be selected to reflect the course and an assessment of how well the children have learnt the material can be made. Traditional tests usually attempt to assess a "generalised subject area" and thus the area of particular interest to this particular teacher would probably only be represented by a few items; and yet all the items need to be attempted to obtain a standardized score. On the other hand, a local authority may be

interested in this "generalised subject area" but because they are interested in the performance of large groups rather than of individual pupils the number of items in the test could be reduced and still give sufficiently accurate information. Item banks would allow a test to be rapidly constructed for a specific purpose and reduce the inefficiency of testing which occurs at present.

Another advantage of item banks is that, since tests are constructed to strict specifications which reflect the interests of the user, the information can be used diagnostically. Since teachers, for example, have such a wide range of views on which aspects of a subject are meaningful, teachable areas, a large number of diagnostic tests constructed on traditional lines would be needed to do the same job.

A problem encountered by the user of traditional tests is that a test becomes inappropriate over time. This can be due to cultural or curriculum changes or simply because the test becomes too easy for a child as he gets older. However, if a new test is selected it is not easy to compare a child's present performance with his previous performance. This problem may be more obvious if we consider a practical situation. A test is given to a young child to assess his attainment in mathematics. Since the child has only just begun to learn elementary set theory, a test is chosen to cover this aspect of mathematics. In a few years time this test will become inappropriate since many more areas of mathematics will have been covered. A new test must therefore be chosen, but the scores on the new test cannot be directly related to those on the previous test. The new score compares the child to his new age group. We do not, therefore, have a way to monitor directly a child's progress over time. Similarly, attempts to monitor group performances over time usually require new tests as the old ones become outdated. The new tests need to be standardized on the present population and thus the scores cannot easily be related to the scores obtained on the previous test.

This problem is overcome to some extent by using the measurement scale described earlier. Since the items in a test can be changed in order to make the test more relevant to the people taking it, and yet the scores can still be translated into an ability or attainment estimate on this scale, the problem of monitoring is much easier. There are still problems however which remain to be sorted out by empirical investigation. For instance

- (a) no one would suggest that a unique hierarchy of difficulty exists across different subjects. Likewise we may find that this

- hierarchy of difficulty does not even exist within subject areas;
- (b) the hierarchy of difficulty could be upset by different degrees of exposure to a subject. Intensive coaching could alter the items' relative difficulties or the fact that certain members of our calibration sample were never even taught the subject could produce unstable difficulty estimates;
 - (c) the use of a single scale of achievement could encourage the teaching of a narrow area of the curriculum and the selection of items to represent this area. Obviously scale scores thus obtained would not be generalizable to the whole subject area.

The development of modular type testing so that assessment becomes descriptive as well as quantitative could help solve these problems. Modules could be developed for areas where the hierarchy of difficulty is stable. Teaching a narrow area of the curriculum would then be reflected in the narrow choice of modules.

The advantage of an item bank system is that these problems, already present with traditional tests, can begin to be investigated. A model such as the Rasch Model can be used to investigate the question of when meaningful comparisons can be made. We can begin to make more informed judgements about whether to monitor progress on a single scale and we can begin to look at issues relating to the way that a subject is learnt or taught. Traditional tests have relied too heavily for their construction on subjective judgements, hallowed tradition and out-of-date research and results from such tests have not been sufficiently generalizable to answer some of the important questions expected of them. Item banks provide a vehicle for more sophisticated inferences to be made from test results.

In Conclusion

As mentioned earlier the advantages of item banks would appear to hinge on their flexibility while providing scores on a single scale. The flexibility means that objective tests can be constructed to fit varying curricula while still relating the levels reached in the different areas. The tests constructed can be more descriptive since they are chosen to reflect well-defined subject areas rather than a generalised curriculum which is necessarily the case with most traditional tests. They can be used for a larger variety of purposes such as diagnosis or monitoring and the test can reflect these purposes more easily since, for example, shorter tests can be constructed for the purpose of monitoring group performances.

This leaflet may give the impression that item banks are simple, straightforward systems to develop. The fact that they have only been developed on a limited scale (especially in Britain) is an indication that this is not so. There are considerable problems still to investigate. The simple model described here for obtaining a measurement scale, for instance, may still prove inadequate when banks are developed on a large scale. However, preliminary work indicates that the ideas developed here are certainly very feasible. As mentioned earlier, NFER has developed an experimental item bank in mathematics for 2nd, 3rd and 4th year secondary. The bank is available in booklet form from which teachers select their own items and construct their own tests. Since it is still an experimental version we would welcome it if people could return to the NFER Test Department any data obtained.

The Assessment and Performance Unit set up by the DES to monitor standards is also committed to item banking although they are not yet committed to the scale of measurement described here. In addition a project has been set up at the NFER to provide item bank facilities in language and maths to LEAs and possibly to teachers. It is hoped that a marketable product will be available before the end of this decade. Many problems, however, remain unsolved but the possibilities of item banks are too great to let pass. It is likely that they will form a significant part of the field of assessment in one form or another in the very near future.

Suggested Readings

- (1) Childs, Roy *Norm Referenced Testing and the Standard Scores*. NFER, 1977 (leaflet).
- (2) Choppin, Bruce *Item Banking and the Monitoring of Achievement: An Introductory Paper*. NFER, 1978.
- (3) Whitely, Susan and Davies, Rene: 'The nature of objectivity with the Rasch Model', *Journal of Educational Measurement*, 1974, Vol II, No. 2, pp. 163-178.
- (4) Wood, R. and Skurnik, L.S. *Item Banking*. NFER, 1969.